Summarizing Measured Data

Kai Shen

Dept. of Computer Science, University of Rochester

Summarizing Measured Data

- Large number of measurements/observations on a random variable/metric
  - request arrival inteval, CPU utilization, speedup, …
- Summarizing them for easy understanding, comparison, reconstruct …
- Summarizing:
  - to a single representative
  - variability
  - distribution

To A Single Representative

- Mean – average
- Median – middle value in ranked order
- Mode – most concentrated value
  - vague definition for samples with continuous values
- Which one to use? Criteria:
  - Impact of outlier?
  - Is total of interest?
  - Deviation to the representative
    - minimize the expected deviation: E(|x-x*|)
    - high probability that the deviation is very small

Simple Extensions

- Constant time scaling of the random variable
  - mean, median, and mode all scale
- Sum of two random variables
  - sum of the two means to be the new mean
  - don't know about median or mode
- Product of two random variables
  - product of the two means to the new mean?

Other Means

- Geometric mean
  - $x^* = (x_1 . x_2 \ldots x_n)^{1/n}$
  - Good when the product of all values is meaningful
  - $gm(x_1/y_1, \ldots, x_n/y_n) = gm(x_1, \ldots, x_n) / gm(y_1, \ldots, y_n)$
  - Also good for representing mean of ratios
    - not necessarily the case – mean of CPU utilizations
- Harmonic mean
  - $X^* = n / (1/x_1 + 1/x_2 + \ldots + 1/x_n)$
  - Good then the arithmetic mean of the inverse is meaningful
    - instruction per cycle vs. runtime

Summarizing Variability

Then there is the man who drowned crossing a stream with an average depth of six inches.

- W. I. E. Gates

- Variability, or index of dispersion
  - range – min/max values
  - variance or standard deviation
  - x-percentile (e.g., 10 and 90-percentile)
  - mean absolute deviation: $E(|x-x^*|)$

Selecting Index of Dispersion

- Impact of outliers
  - range > variance > mean absolute deviation > percentile ≈ 0
- Calculation costs
- Characterizing the network traffic

Determining Data Distributions

- Draw histograms and then visual inspection
- Related to regression

Disclaimer

- Most materials in these slides were developed from the book "The Art of Computer Systems Performance Analysis", R. Jain, 1991, Wiley.